GC content and recombination: reassessing the causal effects for the Saccharomyces cerevisiae genome.

Marie-Claude Marsolier-Kergoat, Edouard Yeramian

Genetics: Published Articles Ahead of Print, published on July 13, 2009 as 10.1534/genetics.109.105049

GC content and recombination: reassessing the causal effects for the *Saccharomyces cerevisiae* genome

Marie-Claude Marsolier-Kergoat[*] and Edouard Yeramian[†]

[*]Institut de Biologie et de Technologies de Saclay, CEA/Saclay, 91191 Gif-sur-Yvette, France

[†]Unité de Bio-Informatique Structurale, CNRS URA 2185, Institut Pasteur, 75724 Paris, France

Running title: GC and recombination: causality in yeast

Keywords: recombination, *Saccharomyces cerevisiae*, GC content, biased gene conversion

Corresponding author:

Marie-Claude Marsolier-Kergoat

Institut de Biologie et de Technologies de Saclay, Service de Biologie Intégrative et de Génétique Moléculaire, Bât. 144, CEA/Saclay, 91191 Gif-sur-Yvette Cedex, France

Phone: 33-1-69-08-83-54

Fax: 33-1-69-08-47-12

E-mail: mcmk@cea.fr

ABSTRACT

Recombination plays a crucial role in the evolution of genomes. Amongst many chromosomal features, GC content is one of the most prominent variables that appear to be highly correlated with recombination. However, it is not yet clear 1) whether recombination drives GC content (as proposed for example in the biased gene conversion model), or the converse and 2) what are the length scales for mutual influences between GC content and recombination. Here we have reassessed these questions for the model genome *Saccharomyces cerevisiae*, for which the most refined recombination data are available. First, we confirmed a strong correlation between recombination rate and GC content at local scales (a few kilobases). Second, based on alignments between *S. cerevisiae*, *S. paradoxus* and *S. mikatae* sequences, we showed that the inferred AT/GC substitution patterns are not correlated with recombination, indicating that GC content is not driven by recombination in yeast. These results thus suggest that, in *S. cerevisiae*, recombination is determined either by the GC content or by a third parameter, also affecting the GC content. Third, we observed long-range correlations between GC and recombination for the chromosome III (for which such correlations were reported experimentally and the model for many structural studies). However similar correlations were not detected in the other chromosomes, restraining thus the generality of the phenomenon. These results pave the way for further analyses aiming at the detailed untangling of drives involved in the evolutionary shaping of the yeast genome.

INTRODUCTION

The architecture of genomes is the result of various evolutionary forces, which can exert concerted or opposing effects. Recombination is considered to represent one such fundamental drive. Indeed correlations with recombination were reported for a large number of structural or functional properties, such as the length of genes, the length of introns for split genes (COMERON and KREITMAN 2000; PRACHUMWAT *et al.* 2004) or even gene order, with the clustering of essential genes in regions of low recombination (PAL and HURST 2003). GC content represents perhaps the most prominent property for which strong correlations with recombination were reported for the genomes of many organisms including mammals, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae* (BIRDSELL 2002; GERTON *et al.* 2000; KONG *et al.* 2002; MARAIS *et al.* 2001; MEUNIER and DURET 2004). On the other hand it was recently demonstrated that in *Arabidopsis thaliana* rate of crossover and GC content are not correlated (DROUAUD *et al.* 2006). However, despite these numerous results, it is not clear as yet 1) whether recombination drives GC content or the converse and 2) what are the length scales for the correlations between GC and recombination.

Correlations between recombination and GC content have been detected both at local scales (typically in the kilobase (kb) range, see (GERTON *et al.* 2000)) and at much larger ones (KONG *et al.* 2002). Arguments were advanced in favor of context-dependent recombinational activities, with the idea that such activities could be regulated, at least in part, by global features of chromosome structure, characterized more or less directly by the GC content (for a general overview, see for example (EYRE-WALKER and HURST 2001)). In this direction, in terms of evolutionary models, mutual influences between recombination and GC were even considered at the highest organizational levels, with the proposal that the large-scale

organization of mammalian genomes in terms of GC-rich isochores could be accounted for to a large extent by the integral of past recombinational activities (DURET and ARNDT 2008; DURET *et al.* 2006).

Regarding the causality relationship between recombination and GC, the biased gene conversion model (see (EYRE-WALKER 1993) for original formulations) proposes that recombination represents a driving force for GC variations, from local to genome-wide scales (in terms of isochore structures). In this model a basic role is attributed to allelic gene conversions during meiotic recombination, as a consequence of the repair of mismatches in heteroduplex DNA. This process is supposed to be biased toward GC, leading to an increase of overall GC contents in regions with high recombination activity (BIRDSELL 2002; BROWN and JIRICNY 1989; EYRE-WALKER 1993; GALTIER *et al.* 2001; MARAIS *et al.* 2001). On the contrary, with analyses mainly based on the *S. cerevisiae* genome, the tenants of the opposite causality model have suggested that it is rather high GC content that promotes recombination (BLAT *et al.* 2002; GERTON *et al.* 2000; PETES 2001; PETES and MERKER 2002).

In this general background we want here to reassess various questions concerning the relationships between recombination and GC for the *S. cerevisiae* model system. Surprisingly, whereas *S. cerevisiae* has served as the system of choice for many of the original questions and models concerning recombination, it appears that various questions, debated notably in the context of mammalian genomes, were not further put to test in the *S. cerevisiae* genome for which the most accurate recombination data in any system have become recently available (BLITZBLAU *et al.* 2007; BUHLER *et al.* 2007; MANCERA *et al.* 2008).

We first addressed the causality question at local scales, using the same approach as the one that was implemented in the case of mammalian genomes. At such scales, with the new recombination data for *S. cerevisiae*, we confirmed the strong correlations between GC and recombination. We then analyzed the patterns of substitutions that occurred in the *S.*

cerevisiae strain S288C lineage under two evolutionary perspectives: 1) after the divergence between the S288C lineage and the lineage of another strain of *S. cerevisiae*, YJM789, and 2) after the divergence between the *S. cerevisiae* and *S. paradoxus* lineages. The rationale behind such substitution analyses (DURET and ARNDT 2008; KHELIFI *et al.* 2006; MEUNIER and DURET 2004; WEBSTER *et al.* 2005) is to address the possible effect of recombination on GC content, through the determination of the relative rates of AT to GC and GC to AT substitutions. Based on such analyses, we found that recombination is not directly correlated to the patterns of AT/GC substitutions in *S. cerevisiae*, which indicates that recombination has no detectable influence on GC content in this case.

Beyond the local scales, we then considered the ranges of mutual influences between recombination and GC content in *S. cerevisiae*. We first extended the substitution analyses at significantly larger scales, in order to test the possibility that the local result could hide long-range correlations. Indeed, results demonstrating the effect of recombination on GC content in the human genome could be observed only at the megabase (Mb) scale (DURET and ARNDT 2008). In *S. cerevisiae* however, we found no evidence for a significant effect of recombination on GC content at any scale. Concerning the large-scale influences, we tested then a model developed by Petes and Merker (PETES and MERKER 2002), following which, in *S. cerevisiae*, recombinational activity at one given locus could be determined by the GC content of the surrounding region, over large distances. This model was elaborated based on the analysis of chromosome III, but our results did not allow to validate the generality of the hypothesis for all *S. cerevisiae* chromosomes.

METHODS

The sequences of *Saccharomyces cerevisiae* strain S288C, *Saccharomyces paradoxus* and *Saccharomyces mikatae* were downloaded from the Broad Institute web page (http://www.broad.mit.edu/annotation/fungi/comp_yeasts/downloads.html; in correspondence with supplementary information in (KELLIS *et al.* 2003)). The sequences of *Saccharomyces cerevisiae* strain YJM789 were retrieved from the Resources for Fungal Comparative Genomics web site (http://fungal.genome.duke.edu/annotations/scer_yjm789/). Multiple sequence alignments were performed using ClustalW (downloaded from http://www.ebi.ac.uk/Tools/clustalw/).

For the substitution analyses of *S. cerevisiae* strains S288C and YJM789 (with *S. paradoxus* as outgroup), we selected all the open reading frames (ORFs) with unambiguous correspondence in *S. cerevisiae* and *S. paradoxus* (listed by Kellis and collaborators in the web page ftp://ftp-genome.wi.mit.edu/pub/annotation/fungi/comp_yeasts/S1b.ORFs/listing.txt). We further restricted our analyses to the ORFs annotated as 'verified' or 'uncharacterized' in the *Saccharomyces* Genome Database (http://www.yeastgenome.org/). A total of 3,997 protein-coding sequences were thus analyzed, along with their intergenic 5' sequences (denoting as intergenes the sequences located between two consecutive verified or uncharacterized ORFs).

For the substitution analyses of *S. cerevisiae* and *S. paradoxus* (with *S. mikatae* as outgroup), we used the sequence of *S. cerevisiae* strain S288C. As above, we selected all the ORFs with unambiguous correspondence in *S. cerevisiae*, *S. paradoxus* and *S. mikatae* (ftp://ftp-genome.wi.mit.edu/pub/annotation/fungi/comp_yeasts/S1b.ORFs/listing.txt). We then restricted our analyses to the ORFs annotated as 'verified' or 'uncharacterized' in the

Saccharomyces Genome Database. A total of 4,295 protein-coding sequences and their intergenic 5' sequences were thus analyzed.

We estimated the substitutions in *S. cerevisiae* strains S288C and YJM789 using *S. paradoxus* as an outgroup to infer the ancestral nucleotide sequences, using parsimony. We disregarded positions where the sequence of *S. paradoxus* differed from the sequences of both *S. cerevisiae* strains. Similarly, we inferred the substitutions in *S. cerevisiae* and *S. paradoxus* lineages using parsimony on informative sites with *S. mikatae* as an outgroup, and we disregarded the sites where *S. mikatae* sequences differed from the sequences of *S. cerevisiae* and *S. paradoxus*. We did not attempt, in neither case, to correct for multiple substitutions. The substitution rates were estimated by dividing the number of inferred substitutions by the number of inferred, potentially mutable, ancestral sites. We calculated the equilibrium GC contents (GC*) using the model of Sueoka (SUEOKA 1962), as the ratio $u/(u+v)$ where u and v represent, respectively, the AT to GC and GC to AT substitution rates (this model assumes that all sites within a sequence evolve independently of one another).

Estimates of recombination rates were from the study of Buhler and collaborators (BUHLER *et al.* 2007), unless otherwise specified. In this study the single-stranded DNA intermediates (ssDNA) resulting from the processing of meiotic double-strand breaks (DSBs) were detected by microarray hybridization in *S. cerevisiae* SK1 strain. We took as an estimate of the recombination rate for a given sequence (protein-coding sequence or intergene) the average of the measured values for DNA probes with midpoints localized between the start and end points of the sequence (average ratios of background-normalized fluorescence in *dmc1Δ* mutants). It is worth noting also the study mapping meiotic ssDNA by Blitzblau and collaborators (BLITZBLAU *et al.* 2007), concomitant with that of Buhler *et al*. Both studies relied on the same strategy for the SK1 strain, using the same microarrays. In preliminary analyses, we obtained very similar results using the data from Blitzblau *et al.* and from Buhler

et al., as they are highly correlated. Accordingly, in what follows, we present only the results obtained with Buhler *et al.*'s data. Since the recombination rates in the SK1 strain could differ from those in the S288C and YJM789 strains, we also analyzed the recombination rates obtained by Mancera and collaborators by genotyping ~ 52,000 markers in spores derived from 51 meioses of an S288C/YJM789 hybrid strain (MANCERA *et al.* 2008). In this case we considered the adjusted counts of recombination events including both crossovers and non-crossovers.

We analyzed the correlations between recombination rates and GC, or GC*, in non-overlapping DNA windows of variable size (from 5 to 100 kb) by pooling the values for protein-coding sequences with start and end points located within the limits of the corresponding windows. The GC value associated with a given window was estimated by dividing the total number of Gs and Cs in the protein-coding sequences within the window, by the sum of the lengths of these sequences. Similarly, the GC* value associated with a given window was calculated from the global AT to GC and GC to AT substitution rates for the protein-coding sequences in the window (estimated by dividing the total number of AT to GC or GC to AT substitutions by the total number of potentially mutable, ancestral sites). Similarly, the recombination rate associated with a given window was taken as the average of the recombination rates estimated for the protein-coding sequences within the window, weighted by the lengths of these sequences.

Data sets were produced and analyzed with custom Python scripts (available upon request). Statistical analyses were performed with the R environment (R DEVELOPMENT CORE TEAM 2008). Given the non-normality of the distributions of several variables (in particular recombination rates), we generally resorted to the correlation of Spearman (coefficient ρ) rather than that of Pearson (coefficient r), with the values of the two coefficients being usually notably close.

RESULTS

No detectable influence of recombination on AT/GC substitution patterns at local scales:

We took advantage of the recently published recombination data for *S. cerevisiae* to first reassess the correlations between recombination rate and GC content at local scales. As the bulk of the *S. cerevisiae* genome consists of protein-coding sequences (with short intergenic regions), we first focused on the protein-coding sequences of *S. cerevisiae* strains S288C and YJM789. The average divergence in coding sequences between these two strains amounts to ~ 0.4 % (Table 1). With substitution analyses of S288C and YJM789, using *S. paradoxus* as outgroup, we inferred the occurrence of 11,101 AT to GC substitutions and 9,281 GC to AT substitutions in the two strain lineages, since their divergence (see Methods). As the substitution rates in S288C and YJM789 lineages appeared to be weakly correlated (ρ = 0.39, $P < 10^{-10}$ for the AT to GC substitution rates; ρ = 0.35, $P < 10^{-10}$ for the GC to AT substitution rates), we avoided to pool together the substitution events inferred in the two lineages. The differences between the relative rates of AT to GC and GC to AT substitutions were quantified for each gene based on the calculation of the equilibrium GC contents (GC*), the GC contents that would be reached at equilibrium by the sequences if patterns of substitutions remained unchanged (see Methods). We found that the GC* values for S288C and YJM789 genes are not correlated (ρ = 0.012, P = 0.56, N = 2314).

The recombination rate associated with each protein-coding sequence was estimated from the ssDNA abundance measurements in the study of Buhler and collaborators (BUHLER *et al.* 2007). In agreement with previous studies (BIRDSELL 2002), we found that GC content and recombination rate are strongly correlated for *S. cerevisiae* genes (ρ = 0.54, $P < 10^{-10}$) (Table 1). This correlation value appears to be higher than that reported previously (ρ = 0.33) (BIRDSELL 2002), possibly because of the enhanced accuracy in the measured recombination

rates. Indeed, the recombination data used by Birdsell had been measured in *rad50S* cells (GERTON *et al.* 2000), in which the distributions of meiotic DSBs can be significantly different from those in the wild type (BUHLER *et al.* 2007). As another way to estimate the correlation between recombination and GC content, we analyzed the GC contents of the 5 kb regions around the 1179 recombination hotspots determined by Buhler and collaborators (corresponding to a threshold value of 5 times that of the background, (BUHLER *et al.* 2007)). We found that 73% of these hotspot regions display a GC content higher than the median GC content of 5 kb regions over the genome ($P < 10^{-10}$).

If recombination were driving GC content in yeast, recombination rates should correlate more strongly with GC* (reflecting the recent patterns of AT/GC substitutions), than with the present GC contents of sequences (reflecting the successive substitution patterns throughout their evolutionary history). We observed no correlation between recombination rate and GC* (S288C: ρ = -0.017, P = 0.38, N = 2,779; YJM789: ρ = 0.011, P = 0.54, N = 2,753) (Table 1). We reiterated this analysis taking into account codon positions (Table 1). We again found no significant correlation ($P < 0.01$) between recombination rates and GC*, except for the second codon position in S288C (weak and negative correlation). The coefficients for the correlation between GC content and recombination were different for the three codon positions, with the highest correlation being for the third codon position.

DSBs occur mostly in intergenic promoter-containing regions (BAUDAT and NICOLAS 1997; GERTON *et al.* 2000). Accordingly we further analyzed intergenic sequences in relation with recombination, both at bulk level and following promoter-related properties. The bulk analyses (Table 1) led to similar conclusions as those above. Regarding promoter analysis, we defined 4,241 groups composed of intergenic sequences containing at least one promoter region, along with the two adjacent coding sequences. Indeed, according to current DSB repair models, upon initiation in intergenic promoter-containing regions, ssDNA formation

propagates bidirectionally towards the two adjacent coding sequences. The analysis of correlations between the averaged recombination rate, GC* and GC content for the 4,241 groups revealed again a strong correlation between GC and recombination rate and lack of correlation between recombination rate and GC* (Table 1).

Next we examined the correlation between GC* and recombination rate by analyzing the pattern of substitutions that occurred in *S. cerevisiae* strain S288C lineage after the divergence between *S. cerevisiae* and *S. paradoxus*. The average divergence in coding sequences between *S. cerevisiae* strain S288C and *S. paradoxus* amounts to ~ 10.3 % (Table 2), and accordingly the number of inferred substitutions that could be analyzed was larger than in the previous case (160,995 AT to GC substitutions and 132,058 GC to AT substitutions in *S. cerevisiae* lineage, and 129,907 AT to GC substitutions and 85,697 GC to AT substitutions in *S. paradoxus* lineage). For the whole coding sequences (all codon positions), we observed a weaker correlation between recombination rate and GC* ($\rho = 0.092$, $P = 2.10^{-9}$, $N = 4209$) than between recombination rate and GC ($\rho = 0.54$, $P < 10^{-10}$) (Table 2). In addition, we found that the direct correlation between recombination rate and GC* (as estimated by the partial correlation coefficient) is small and negative ($r = -0.063$), indicating that the correlation between recombination rate and GC* results from the correlation between GC and GC* ($\rho = 0.30$, $P < 10^{-10}$). The correlation between GC and GC* is attributable to the high number of substitutions that *S. cerevisiae* genes have experienced since the divergence between *S. cerevisiae* and *S. paradoxus*, with a direct impact on the present GC content.

Similar results were obtained for the intergenic sequences as well as in analyses taking into account the codon positions. In all cases, the correlation between recombination rate and GC appeared stronger than the correlation between recombination rate and GC*, this latter correlation being attributable to the correlation between GC and GC*, as demonstrated by partial correlation analyses (Table 2; data not shown).

Finally we reiterated the various analyses above with the recombination data from Mancera *et al.* (MANCERA *et al.* 2008) (see Methods) and obtained similar results (Tables S1 and S2). However the correlations between recombination and GC content obtained with the data from Buhler *et al.* (BUHLER *et al.* 2007) were systematically higher than those with the data from Mancera *et al.* This difference probably comes from the fact that Buhler *et al.*'s data correspond to values averaged over millions of cells whereas Mancera *et al.*'s data are based on the analysis of 204 cells. In our subsequent analyses we used the recombination data from Buhler *et al.*

To summarize, following the various analyses above, we found no evidence in *S. cerevisiae* for any direct influence of local recombination rates on local AT/GC substitution patterns. Accordingly, the observed correlation between local recombination rate and local GC content in yeast cannot be accounted for by a causal influence of recombination on GC content.

No detectable influence of recombination on AT/GC substitution patterns at regional scales:

The conclusions above can appear to contradict those reported for the human genome, in which the correlation between crossover rate and GC* is stronger than that between crossover rate and GC content (DURET and ARNDT 2008; MEUNIER and DURET 2004). For the human genome it was accordingly concluded that the correlation between GC content and crossover rate is primarily attributable to the influence of recombination on AT/GC substitution patterns. In this genome, the reported strong correlations between crossover rate and GC* are however observed at the Mb scale. The strength of this correlation decreases significantly with window size, becoming very weak below 200 kb (DURET and ARNDT 2008). Following these observations we could ask the question of the possible existence of a

correlation between recombination rate and GC* in yeast at larger scales, despite the absence of such a correlation at small scales.

We therefore analyzed the correlations between recombination rate and GC, or GC*, in *S. cerevisiae* strain S288C at large scales by pooling genes located in DNA windows ranging from 5 to 100 kb (see Methods). We considered the protein-coding sequences, taking into account all codon positions. We found that the correlation between recombination rate and GC content decreases with increasing window sizes (Figure 1). On the other hand the correlation between recombination rate and GC* appears to fluctuate, with values systematically lower than those of the correlation between recombination rate and GC. Moreover, we found that the weak correlation observed between recombination rate and the GC* (determined from the substitutions having occurred in S288C lineage after the divergence between *S. cerevisiae* and *S. paradoxus* (Figure 1B)) is in this case also directly attributable to the correlation between GC and GC* (as determined by partial correlation analysis, data not shown). Similar results were obtained with the strain YJM789 used as reference (Figure S1).

In conclusion, as a difference with the human genome, we found no significant correlation in yeast between recombination rates and AT/GC substitution patterns for length scales ranging from 1 to 100 kb. There is accordingly no evidence in *S. cerevisiae* for a notable influence of recombination on GC content.

As for the correlation between recombination rate and GC, we observed that its strength decreases with increasing scales (Figure 1), indicative of a correlation which is mainly at local scales (a few kb). However, this predominant effect did not rule out the possibility for local recombination rate to be under the influence of regional GC content. We examined next such a possibility.

Correlation between local recombination rate and regional GC content:

Several authors have proposed that recombination and GC content could be correlated not only at local, but also at regional (>30 kb) scales (BLAT *et al.* 2002; PETES and MERKER 2002). Petes and Merker based their conclusion on the re-analysis of data from Borde and collaborators, who had measured both meiotic recombination rates and the formation of DSBs in recombination reporter constructs, inserted at ten locations in the chromosome III (BORDE *et al.* 1999). In their work, Borde *et al.* demonstrated that the recombination activity of the constructs reflected the recombination activity of the loci in which they were inserted. It was then concluded that the recombination activity of the inserts was governed by their chromosomal context. Petes and Merker found that the recombination activity of the inserts was strongly correlated with the global GC content (with no distinction made between protein-coding sequences and intergenes) of the chromosomal sequences flanking the insertion, measured within DNA windows ranging from 0.5 to 100 kb (PETES and MERKER 2002).

Insertion of a unique reporter construct at different loci allows the elimination of local GC content as a variable, considering that the recombination rate of a locus could be influenced both by its own GC content and by the regional GC content of its chromosomal location. In such context, an influence of the regional GC content on the recombination rate can be more easily demonstrated. We asked whether the influence of the regional GC content on recombination could be also detected based on the recombination rates of the endogenous sequences.

Using 5 kb, non-overlapping windows, we analyzed the correlation between the average recombination rate of each window (estimated from ssDNA measures (BUHLER *et al.* 2007)) and the GC contents of the 20, 50 and 100 kb regions centered on that window (hereafter called GC20, GC50 and GC100, respectively), for all *S. cerevisiae* chromosomes.

In the definition of the 20, 50 and 100 kb regions, we excluded the 5 kb central sequence, to eliminate the contribution of the local GC content. As shown in Table 3, the majority of the chromosomes displayed no significant correlation between recombination rate and GC20, GC50 and GC100 ($P > 0.01$). A significant positive correlation between recombination rates and GC20, GC50 and GC100 was only observed for the chromosome III, in agreement with the results of Petes and Merker (PETES and MERKER 2002). Interestingly, chromosomes XIII and XV displayed a significant negative correlation between both GC50 and GC100 and recombination rates. As a matter of comparison, Table 3 shows the correlations between the recombination rates and the GC contents of 5 kb regions (denoted GC5).

In summary, our analyses do not support a general model in yeast, for high GC content stimulating recombination over large scales. They also highlight the peculiar properties of the chromosome III, the sex chromosome in *S. cerevisiae*. It is then all the more worth noting that the chromosome III has been the model object for many studies devoted to recombination (see for example (BAUDAT and NICOLAS 1997; BLAT and KLECKNER 1999; BLAT *et al.* 2002; BORDE *et al.* 1999)). Following our results here, generalizations from analyses restricted to this chromosome should then be considered with caution.

DISCUSSION

GC content not driven by recombination in *S. cerevisiae*, at any scale:

In our study here, we found no evidence for correlation between recombination rate and AT/GC substitution patterns (as reflected by GC*) in *S. cerevisiae*, for scales ranging between a few kb and 100 kb. The recent AT/GC substitution patterns in *S. cerevisiae* strain S288C were analyzed by inferring the substitutions that took place either after the divergence between S288C and YJM789 or after the divergence between *S. cerevisiae* and *S. paradoxus*. In both cases, the results indicate that recombination has no detectable influence on the evolution of GC content in yeast. This conclusion stands in contrast to that reported for mammalian systems, and notably for the human genome (DURET and ARNDT 2008; MEUNIER and DURET 2004). Duret and collaborators have established that recombination drives the evolution of GC content in the human genome at large scales, by showing that crossover rate correlates more strongly with GC* than with GC. In the human genome, the correlation between crossover rate and GC* is strongest at the 10 Mb scale and decreases with decreasing scales to become very weak below 200 kb. Since the lengths of yeast chromosomes range from 0.23 to 1.53 Mb, we could not explore the relationship between recombination and GC* above the 100 kb scale. However, 100 kb for the 13 Mb yeast genome can be put in correspondence with 25 Mb for the 3,300 Mb human genome. In a related way, the average recombination rate per physical distance in yeasts is about 300 times greater than that in humans (PETES 2001), which again can put in correspondence 100 kb of the yeast genome with 30 Mb of the human genome. Following these proportionality arguments, the length scales explored here for the yeast genome should be comparable with those involved in the mammalian studies.

How then could we account for the fact that we detect no correlation between recombination and AT/GC substitution pattern in yeast? Although different mutagenic or selective effects have been hypothesized, biased gene conversion towards GC appears now as one of the most prominent models to explain the effect of recombination on GC content in mammals (DURET and ARNDT 2008). Mancera and collaborators, analyzing ~ 6,300 recombination events in 51 meioses of an S288C/YJM789 hybrid strain, observed a significant 1.4% GC increase in the converted sequences of the spores, relative to the base content at marker positions in the parental genomes (MANCERA *et al.* 2008). This observation indicates that biased gene conversion operates in the course of recombination in *S. cerevisiae*. Two hypotheses at least could account for the fact that we could not detect the effect of biased gene conversion on GC content in yeast. First, its influence could be masked by other, stronger drives. Second, the location of the recombination hotspots could move on rapid time scales. With this respect it is informative to consider the hypotheses put forward to explain the results in mammals (DURET and ARNDT 2008; MYERS *et al.* 2005), with crossover rate correlating with GC* at the Mb scale but not at the 100 kb scale: short-lived hotspots (demonstrated in humans and chimpanzees (PTAK *et al.* 2005; WINCKLER *et al.* 2005)) along with more conserved recombination density at regional scales. Similarly, in yeast, the lack of correlation between recombination and GC* at various scales could be accounted for with a scheme involving rapidly moving recombination hotspots with no conservation of the regional recombination density.

Local recombination rate correlates with local, but not with regional, GC content in *S. cerevisiae*:

Confirming and amplifying previous results (BIRDSELL 2002; GERTON *et al.* 2000), we observed a strong correlation at local scales (a few kb) between recombination rate and GC

content. On the other hand we found no evidence for a general correlation between regional GC content (measured over 20-100 kb) and local recombination rate. Since our results showed clearly that recombination does not correlate with GC*, at any scale, a simple recombination-driven GC-content evolution scheme can be ruled out for *S. cerevisiae*. What could then account for the correlation between GC content and recombination in yeast ? One possibility is that GC content could influence recombination rate. The determinants of the locations of meiotic DSBs remain poorly understood in yeast (LICHTEN 2008) and GC content could affect one or several of them. For example, the recombination machinery could have a higher affinity for GC-rich sequences or GC content could affect histone modifications or nucleosome positioning, making some spots more or less permissive to the formation of DSBs. Alternatively, a third parameter could influence both recombination and GC content.

In conclusion, our results show that in *S. cerevisiae* recombination and GC content are correlated systematically only at local scales and that in *S. cerevisiae* recombination has no detectable influence on the evolution of GC content, in contrast to mammalian systems. This analysis has also revealed that in yeast recombination is driven either by the GC content or by a third parameter affecting also the GC content. This influence on recombination could also be present in mammals but could be masked by the strong influence of recombination on the GC content. With recombination now set aside as a major drive in yeast, the search in this system for the basic determinants of the GC content is still an open question for future investigations.

ACKNOWLEDGMENTS

We thank Claude Thermes and Yves d'Aubenton-Carafa for helpful discussions and Raphaël Guérois for discussions and for his invaluable help in Python. We acknowledge

many insightful comments and suggestions from two anonymous referees. M.-C. M.-K. was partly financed by the Association pour la Recherche sur le Cancer (ARC) and by the Agence Nationale de la Recherche (ANR). Research was also supported by grant DGA/SSA/CO06co006 from the French Army to E. Y.

REFERENCES

BAUDAT, F., and A. NICOLAS, 1997 Clustering of meiotic double-strand breaks on yeast chromosome III. Proc Natl Acad Sci U S A **94:** 5213-5218.

BIRDSELL, J. A., 2002 Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol Biol Evol **19:** 1181-1197.

BLAT, Y., and N. KLECKNER, 1999 Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. Cell **98:** 249-259.

BLAT, Y., R. U. PROTACIO, N. HUNTER and N. KLECKNER, 2002 Physical and functional interactions among basic chromosome organizational features govern early steps of meiotic chiasma formation. Cell **111:** 791-802.

BLITZBLAU, H. G., G. W. BELL, J. RODRIGUEZ, S. P. BELL and A. HOCHWAGEN, 2007 Mapping of meiotic single-stranded DNA reveals double-stranded-break hotspots near centromeres and telomeres. Curr Biol **17:** 2003-2012.

BORDE, V., T. C. WU and M. LICHTEN, 1999 Use of a recombination reporter insert to define meiotic recombination domains on chromosome III of Saccharomyces cerevisiae. Mol Cell Biol **19:** 4832-4842.

BROWN, T. C., and J. JIRICNY, 1989 Repair of base-base mismatches in simian and human cells. Genome **31:** 578-583.

BUHLER, C., V. BORDE and M. LICHTEN, 2007 Mapping meiotic single-strand DNA reveals a new landscape of DNA double-strand breaks in Saccharomyces cerevisiae. PLoS Biol **5:** e324.

COMERON, J. M., and M. KREITMAN, 2000 The correlation between intron length and recombination in drosophila. Dynamic equilibrium between mutational and selective forces. Genetics **156:** 1175-1190.

DROUAUD, J., C. CAMILLERI, P. Y. BOURGUIGNON, A. CANAGUIER, A. BERARD *et al.*, 2006 Variation in crossing-over rates across chromosome 4 of Arabidopsis thaliana reveals the presence of meiotic recombination "hot spots". Genome Res **16:** 106-114.

DURET, L., and P. F. ARNDT, 2008 The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet **4:** e1000071.

DURET, L., A. EYRE-WALKER and N. GALTIER, 2006 A new perspective on isochore evolution. Gene **385:** 71-74.

EYRE-WALKER, A., 1993 Recombination and mammalian genome evolution. Proc Biol Sci **252:** 237-243.

EYRE-WALKER, A., and L. D. HURST, 2001 The evolution of isochores. Nat Rev Genet **2:** 549-555.

GALTIER, N., G. PIGANEAU, D. MOUCHIROUD and L. DURET, 2001 GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics **159:** 907-911.

GERTON, J. L., J. DERISI, R. SHROFF, M. LICHTEN, P. O. BROWN *et al.*, 2000 Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. Proc Natl Acad Sci U S A **97:** 11383-11390.

KELLIS, M., N. PATTERSON, M. ENDRIZZI, B. BIRREN and E. S. LANDER, 2003 Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423:** 241-254.

KHELIFI, A., J. MEUNIER, L. DURET and D. MOUCHIROUD, 2006 GC Content Evolution of the Human and Mouse Genomes: Insights from the Study of Processed Pseudogenes in Regions of Different Recombination Rates. J Mol Evol.

KONG, A., D. F. GUDBJARTSSON, J. SAINZ, G. M. JONSDOTTIR, S. A. GUDJONSSON *et al.*, 2002 A high-resolution recombination map of the human genome. Nat Genet **31:** 241-247.

LICHTEN, M., 2008 Meiotic chromatin: the substrate for recombination initiation, pp. 165-194 in *Recombination and meiosis: models, means and evolution*, edited by R. EGEL and D.-H. LANKENAU. Springer-Verlag, New-York.

MANCERA, E., R. BOURGON, A. BROZZI, W. HUBER and L. M. STEINMETZ, 2008 High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature **454:** 479-485.

MARAIS, G., D. MOUCHIROUD and L. DURET, 2001 Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci U S A **98:** 5688-5692.

MEUNIER, J., and L. DURET, 2004 Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol **21:** 984-990.

MYERS, S., L. BOTTOLO, C. FREEMAN, G. MCVEAN and P. DONNELLY, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310:** 321-324.

PAL, C., and L. D. HURST, 2003 Evidence for co-evolution of gene order and recombination rate. Nat Genet **33:** 392-395.

PETES, T. D., 2001 Meiotic recombination hot spots and cold spots. Nat Rev Genet **2:** 360-369.

PETES, T. D., and J. D. MERKER, 2002 Context dependence of meiotic recombination hotspots in yeast: the relationship between recombination activity of a reporter construct and base composition. Genetics **162:** 2049-2052.

PRACHUMWAT, A., L. DEVINCENTIS and M. F. PALOPOLI, 2004 Intron size correlates positively with recombination rate in Caenorhabditis elegans. Genetics **166:** 1585-1590.

PTAK, S. E., D. A. HINDS, K. KOEHLER, B. NICKEL, N. PATIL *et al.*, 2005 Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet **37:** 429-434.

R DEVELOPMENT CORE TEAM, 2008 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

SUEOKA, N., 1962 On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci U S A **48:** 582-592.

WEBSTER, M. T., N. G. SMITH, L. HULTIN-ROSENBERG, P. F. ARNDT and H. ELLEGREN, 2005 Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. Mol Biol Evol **22:** 1468-1474.

WINCKLER, W., S. R. MYERS, D. J. RICHTER, R. C. ONOFRIO, G. J. MCDONALD *et al.*, 2005 Comparison of fine-scale recombination rates in humans and chimpanzees. Science **308:** 107-111.

FIGURE LEGEND

FIGURE 1. Analysis of the correlations between recombination rate and GC or GC* values, estimated within DNA windows of different sizes. Spearman correlation coefficients ρ for the relationship between recombination and GC content (circles) or between recombination and GC* (squares) are plotted as a function of DNA window size. GC* values were estimated from the substitutions having occurred in *S. cerevisiae* strain S288C lineage either after the divergence between S288C and YJM789 (A) or after the divergence between *S. cerevisiae* and *S. paradoxus* (B).

FIGURE S1. Analysis of the correlations between recombination rate and GC or GC* values, estimated within DNA windows of different sizes. Spearman correlation coefficients ρ for the relationship between recombination and GC content (circles) or between recombination and GC* (squares) are plotted as a function of DNA window size. GC* values were estimated from the substitutions having occurred in *S. cerevisiae* strain YJM789 lineage either after the divergence between S288C and YJM789 (A) or after the divergence between *S. cerevisiae* and *S. paradoxus* (B).

Sequence type	Mean divergence	Recombination vs									
		GC in S288C		GC* in S288C			GC in YJM789		GC* in YJM789		
		ρ	*P*	ρ	*P*	*N*	ρ	*P*	ρ	*P*	*N*
All codon positions	0.0041	0.54	$< 10^{-10}$	-0.017	0.38	2779	0.54	$< 10^{-10}$	0.011	0.54	2753
1st codon position	0.0025	0.24	$< 10^{-10}$	-0.010	0.71	1312	0.20	$< 10^{-10}$	0.048	0.09	1257
2nd codon position	0.0018	0.27	$< 10^{-10}$	-0.112	3.10^{-4}	1021	0.27	$< 10^{-10}$	-0.016	0.61	994
3rd codon position	0.0078	0.56	$< 10^{-10}$	0.016	0.44	2356	0.57	$< 10^{-10}$	0.006	0.76	2346
Intergenes	0.0098	0.48	$< 10^{-10}$	0.021	0.38	1672	0.46	$< 10^{-10}$	0.011	0.65	1640
Groups	0.0046	0.57	$< 10^{-10}$	-0.007	0.75	1942	0.57	$< 10^{-10}$	0.024	0.30	1923

TABLE 1. Spearman correlation coefficients ρ, and associated probabilities of correlation *P*, for the relationship between recombination and the present or the equilibrium GC contents (GC or GC*, respectively). The values are for different sequence types in *S. cerevisiae* S288C or YJM789 strains. *N* corresponds to the number of genes or intergenes used for computing the correlation between recombination and GC or GC* (i.e. the number of genes or intergenes having experienced at least one substitution in a given lineage, AT to GC or GC to AT, and for which the recombination rate could be estimated). Groups are defined by promoter-containing intergenes along with the two adjacent coding sequences. The analyses were performed only for the groups for which the substitution rates could be inferred for both the intergene and the two adjacent coding sequences.

Sequence type	Mean divergence	Recombination vs				
		GC in S288C		GC* in S288C		
		ρ	*P*	ρ	*P*	*N*
All codon positions	0.103	0.54	$< 10^{-10}$	0.092	2.10^{-9}	4209
1st codon position	0.062	0.21	$< 10^{-10}$	0.026	0.10	4034
2nd codon position	0.039	0.22	$< 10^{-10}$	0.022	0.18	3726
3rd codon position	0.207	0.55	$< 10^{-10}$	0.27	$< 10^{-10}$	4171
Intergenes	0.185	0.42	$< 10^{-10}$	0.16	$< 10^{-10}$	3700

TABLE 2. Spearman correlation coefficients ρ, and associated probabilities of correlation *P*, for the relationship between recombination and the present or the equilibrium GC contents (GC or GC*, respectively) in *S. cerevisiae* strain S288C. The GC* values were computed from the inferred substitutions having occurred in the *S. cerevisiae* lineage after the divergence between *S. cerevisiae* and *S. paradoxus*. *N* corresponds to the number of sequences used for computing the correlation between recombination and GC or GC* (i.e. the number of sequences having experienced at least one substitution in *S. cerevisiae* lineage, AT to GC or GC to AT, and for which the recombination rate could be estimated).

Chromosome	rec vs GC20			rec vs GC50			rec vs GC100			rec vs GC5		
	ρ	*P*	*N*	ρ	*P*	*N*	ρ	*P*	*N*	ρ	*P*	*N*
I	0.19	0.23	43	-0.05	0.79	37	-0.01	0.97	27	0.62	9.10^{-6}	45
II	-0.08	0.34	159	-0.08	0.34	153	0.00	0.97	143	0.51	$< 10^{-10}$	161
III	0.43	7.10^{-4}	60	0.37	0.01	54	0.42	0.01	44	0.57	2.10^{-6}	62
IV	0.04	0.47	303	-0.01	0.89	297	-0.07	0.23	287	0.52	$< 10^{-10}$	305
V	0.14	0.14	111	0.21	0.04	105	-0.04	0.70	95	0.47	10^{-7}	113
VI	-0.09	0.55	51	0.05	0.76	45	-0.05	0.76	35	0.49	2.10^{-4}	53
VII	-0.03	0.66	215	-0.10	0.14	209	-0.12	0.10	199	0.45	$< 10^{-10}$	217
VIII	-0.13	0.19	109	-0.15	0.13	103	-0.01	0.94	93	0.51	10^{-8}	111
IX	-0.03	0.79	84	-0.30	0.01	78	-0.25	0.04	68	0.43	3.10^{-5}	86
X	0.02	0.79	146	0.02	0.78	140	-0.21	0.02	130	0.62	$< 10^{-10}$	148
XI	0.12	0.19	130	-0.05	0.58	124	-0.13	0.18	114	0.62	$< 10^{-10}$	132
XII	-0.05	0.51	212	-0.14	0.05	206	-0.18	0.02	196	0.45	$< 10^{-10}$	214
XIII	-0.01	0.89	181	-0.19	0.01	175	-0.19	0.01	165	0.49	$< 10^{-10}$	183
XIV	0.11	0.19	153	0.13	0.12	147	-0.16	0.06	137	0.43	3.10^{-8}	155
XV	-0.10	0.14	215	-0.21	2.10^{-3}	209	-0.24	$7. 10^{-4}$	199	0.50	$< 10^{-10}$	217
XVI	-0.11	0.15	186	-0.18	0.01	180	-0.14	0.07	170	0.53	$< 10^{-10}$	188

TABLE 3. Spearman correlation coefficients ρ, and associated probabilities of correlation *P*, for the relationship between the recombination rates of 5 kb regions and the GC contents of the 20 kb, the 50 kb or the 100 kb windows centered on these regions, with the exclusion of the 5 kb central sequences (rec vs GC20, rec vs GC50, and rec vs GC100, respectively), and for the relationship between the recombination rates and the GC contents of 5 kb regions (rec vs GC5). Significant correlations ($P < 0.01$) are highlighted in grey colour.

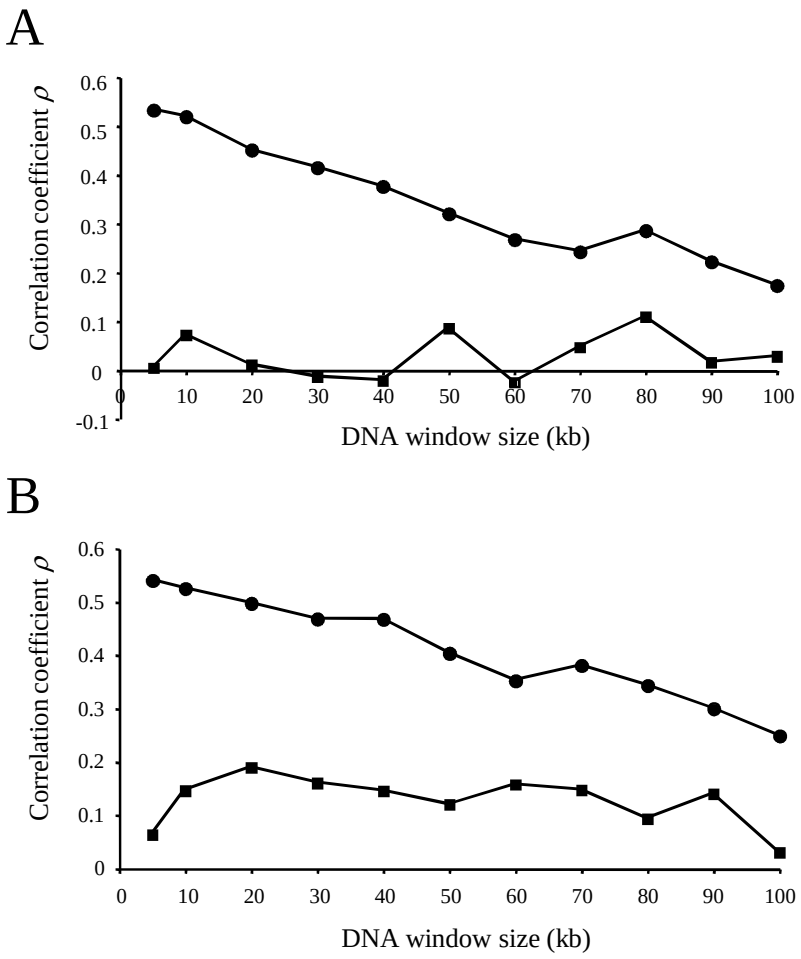
A
Correlation coefficient ρ
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
0
10
20
30
40
50
60
70
80
90
100
DNA window size (kb)
B
Correlation coefficient ρ
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100
DNA window size (kb)

FIGURE 1